BlueGene/L

Hardware Architecture Overview

BlueGene/L design team
IBM Research

BG/L Hardware Architecture - October 2003

- Ruud Haring: BlueGene/L Compute Chip Overview
- Dirk Hoenicke: BLC chip microarchitecture , networks & performance
- Gerry Kopcsay: Power, Packaging, Cooling

Blue Gene/L Partners

Joint Partnership between IBM and Tri-Lab (Lawrence Livermore, Los Alamos, Sandia) ASCI Community.

External Collaborations

- Argonne National Lab
- Barcelona
- Boston University
- Caltech
- Columbia University
- National Center for Atmospheric Research
- Oak Ridge National Lab
- San Diego Supercomputing Center
- Stanford
- Technical University of Vienna
- Trinity College Dublin
- Universidad Politecnica de Valencia
- University of New Mexico
- University of Edinburgh
- University of Maryland

What is BG/L

- A 64k node highly integrated supercomputer based on system-on-a-chip technology.
  - Two ASICs:
    - BlueGene/L Compute (BLC)
    - BlueGene/L Link (BLL)
- Focus is on numerically intensive scientific problems.
- 180-360 TFlop peak performance.
- Strategic partnership with LLNL.
  - Validation and optimization of architecture based on real applications
  - Accustomed to "new architectures" and will work hard to adapt to constraints.
  - Assist us in the investigation of the reach of this machine
- Grand challenge science stress
  - I/O, memory (bandwidth, size and latency), and processing power.

Brief History

- QCDSP (600GF based on Texas Instruments DSP C31)
  - -Gordon Bell Prize for Most Cost Effective Supercomputer in '98
  - -Columbia University Designed and Built
  - -Optimized for Quantum Chromodynamics (QCD)
  - -12,000 50MF Processors
  - -Commodity 2MB DRAM

- QCDOC (20TF based on IBM System-on-a-Chip)
  - -Collaboration between Columbia University and IBM Research
  - -Optimized for QCD
  - -IBM 7SF Technology (ASIC Foundry Technology)
  - -20,000 1GF processors (nominal)
  - -4MB Embedded DRAM + External Commodity DDR/SDR SDRAM

- Blue Gene/L (180/360 TF based on IBM System-on-a-Chip)
  - -Designed by IBM Research in IBM CU-11 Technology
  - -64,000 2.8GF dual processors (nominal)
  - -4MB Embedded DRAM + External Commodity DDR SDRAM

Cost/Performance

- BlueGene/L is cost/performance optimized for a wide class of parallel applications.

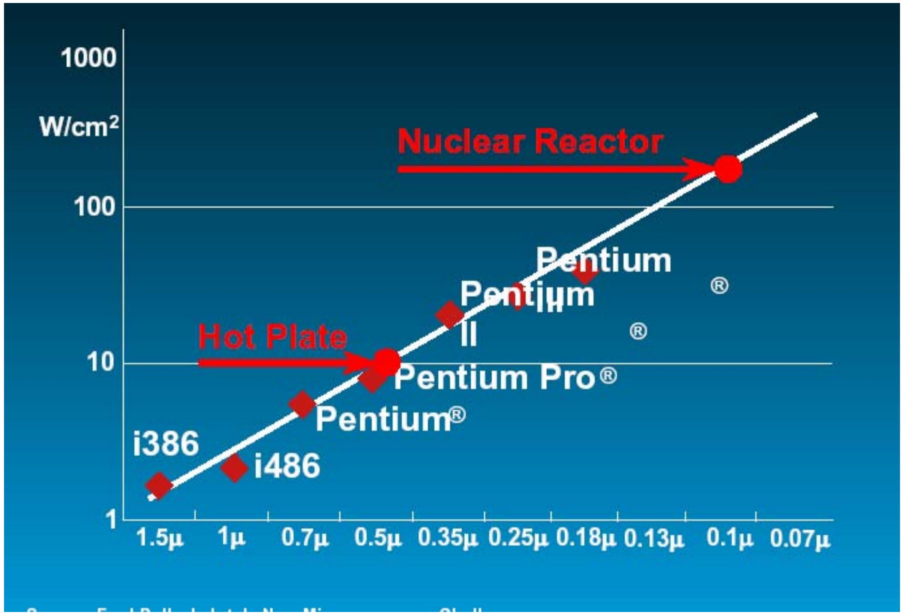
Cost

- Machine
- Facilities
- Hardware Support and Maintenance
- Software Support
  - system
  - application

power is the dominant factor

Performance

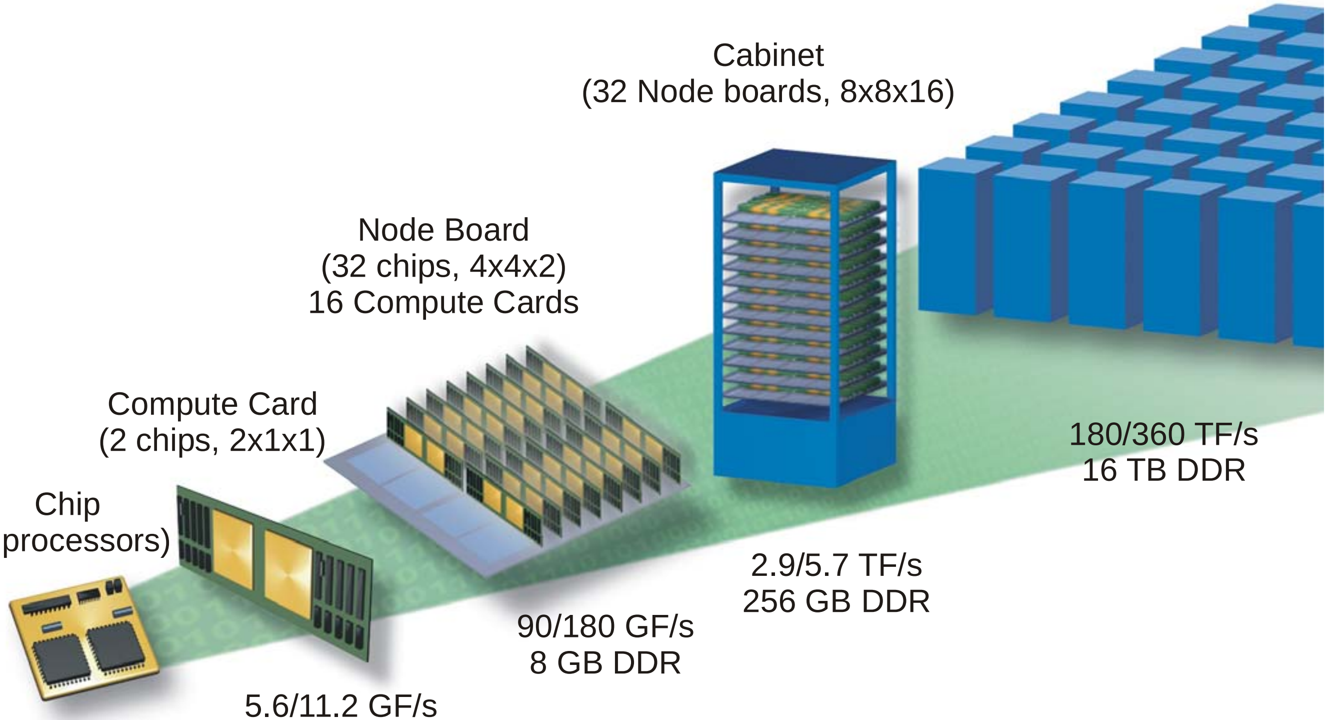
- Peak speed
- Scaleability
- Availability
- Useability
  - tools , debuggers, performance analysis
  - compilers, libraries, frameworks



BG/L Project Motivation

- System on-a-chip offers tremendous cost/performance advantages.
  - Power, Size, Complexity, Design Effort
  - Allows for low latency, high bandwidth memory system
- Scalability of applications to ~100k processors is important research with potentially great payoff.
- Some special purpose machines have had tremendous success using massively parallel.
- Some algorithms are currently scaling to ~thousands of processors

BlueGene/L



System
(64 cabinets, 64x32x32)

Cabinet
(32 Node boards, 8x8x16)

Node Board
(32 chips, 4x4x2)
16 Compute Cards

Compute Card
(2 chips, 2x1x1)

Chip
(2 processors)

180/360 TF/s
16 TB DDR

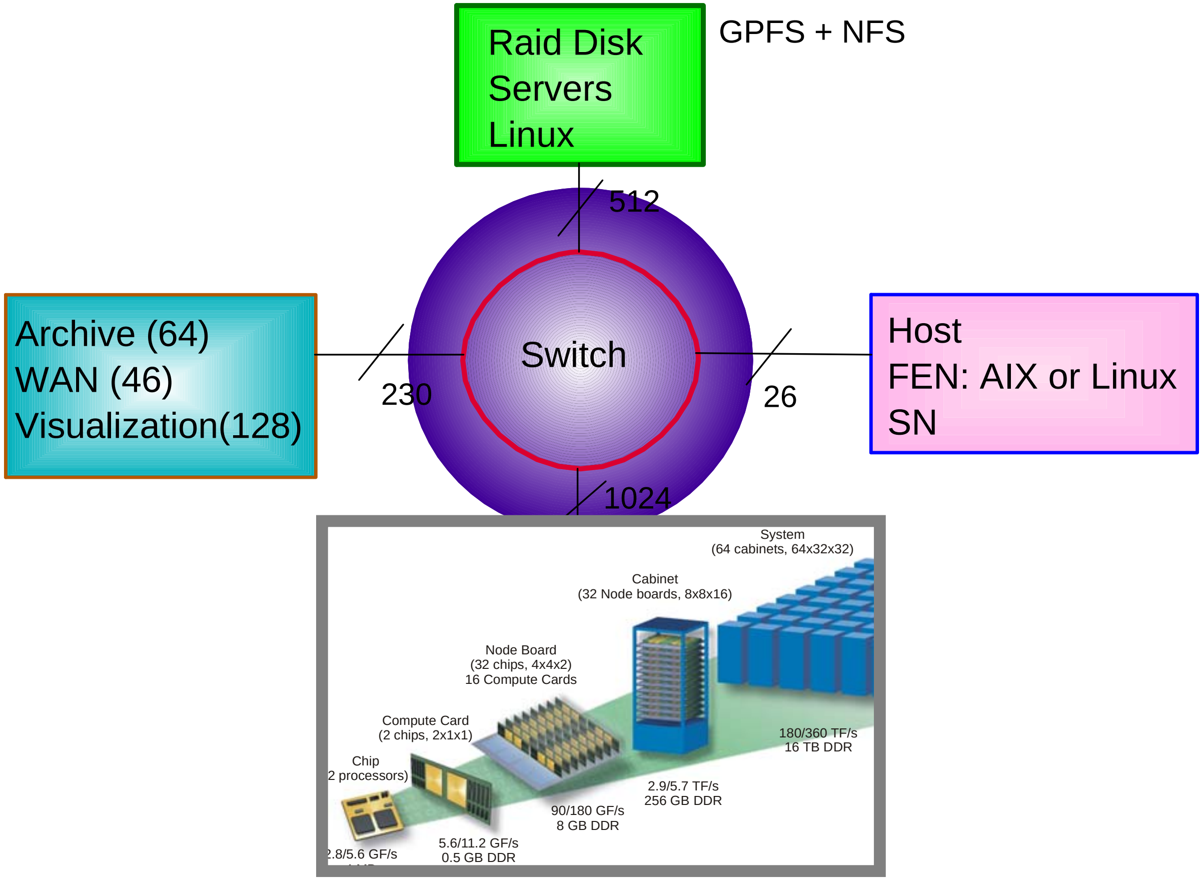
2.9/5.7 TF/s
256 GB DDR

90/180 GF/s
8 GB DDR

5.6/11.2 GF/s
0.5 GB DDR

2.8/5.6 GF/s
4 MB

BlueGene/L System Host



Raid Disk Servers Linux

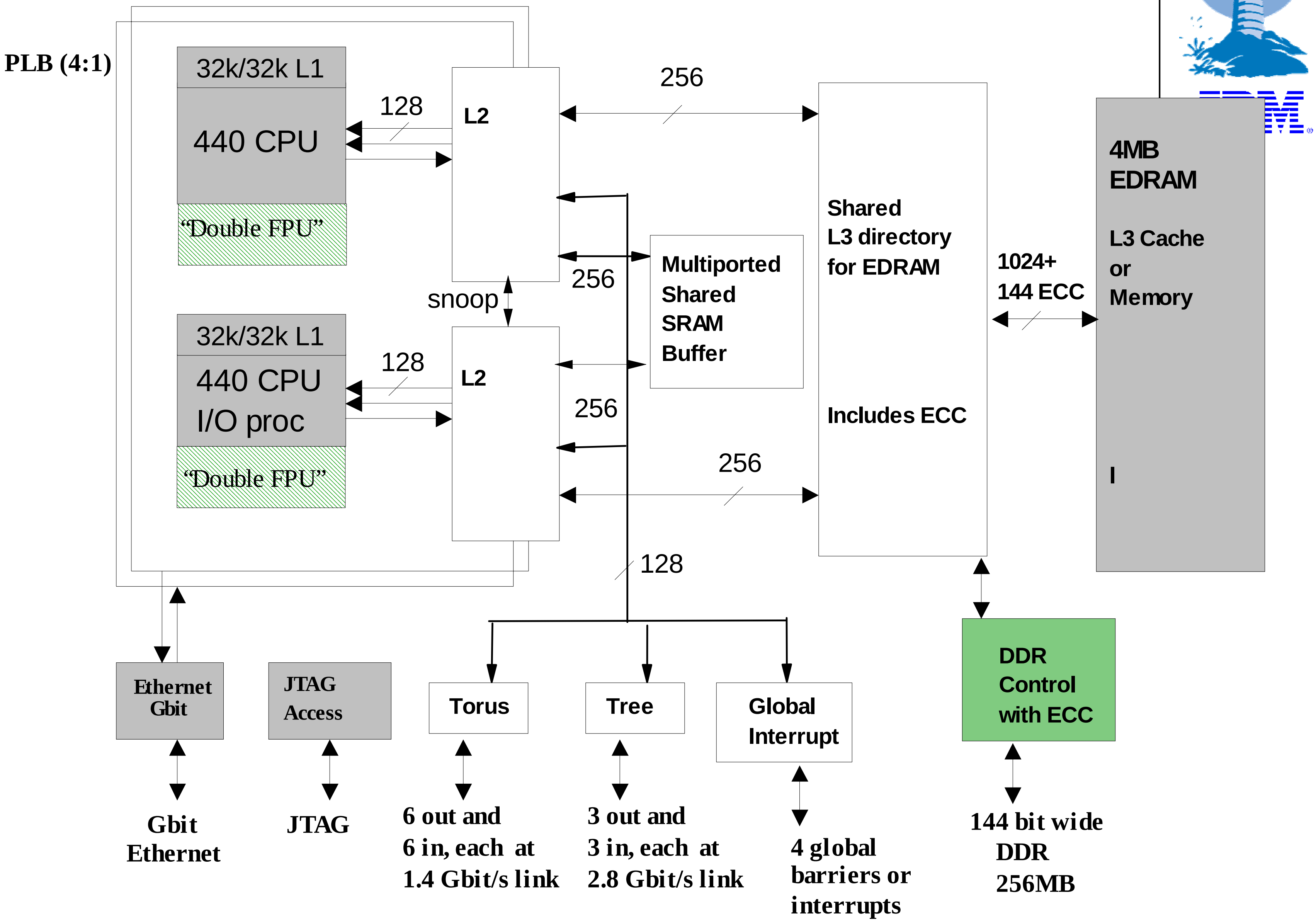
GPFS + NFS

Archive (64)
WAN (46)
Visualization(128)

Switch

Host
FEN: AIX or Linux
SN

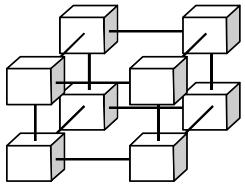
BlueGene/L Compute ASIC



System designed for high reliability

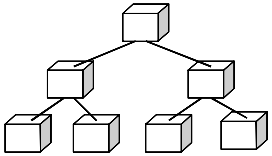
- BLC ASIC
  - All SRAMs in are ECC protected -- except L1 caches in PPC440 and Ethernet
  - L1 caches in 440 cores are parity protected with multiple operating modes
  - Most internal busses have parity detection
  - eDRAM is ECC protected
  - Controller for external DRAM supports memory scrub and ECC with nibble kill reliability. Bit sparing allows for swapping in spare nibble for further reliability.
  - All error types can be counted and used for predictive failure analysis
- Networks:
  - 24 packet CRC + 32 bit “static” CRC
  - Hardware retry for all CRC fails. – Never seen an escape through protocol
  - Optional error injection allows for aggressive testing of link protocol coverage
  - Links are temperature and voltage compensating
- Hardware support for fault isolation
  - Can determine first node that generates a non-repeatable computation in a deterministic calculation
- Redundancy in power, cooling and cabling

The BG/L Networks



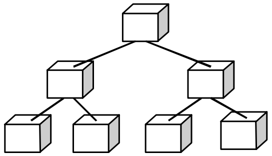
3 Dimensional Torus

- Point-to-point



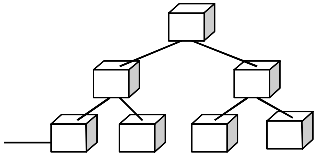
Global Tree

- Global Operations



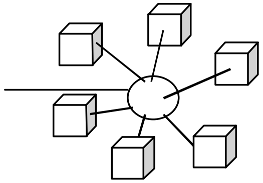
Global Barriers and Interrupts

- Low Latency Barriers and Interrupts



Gbit Ethernet

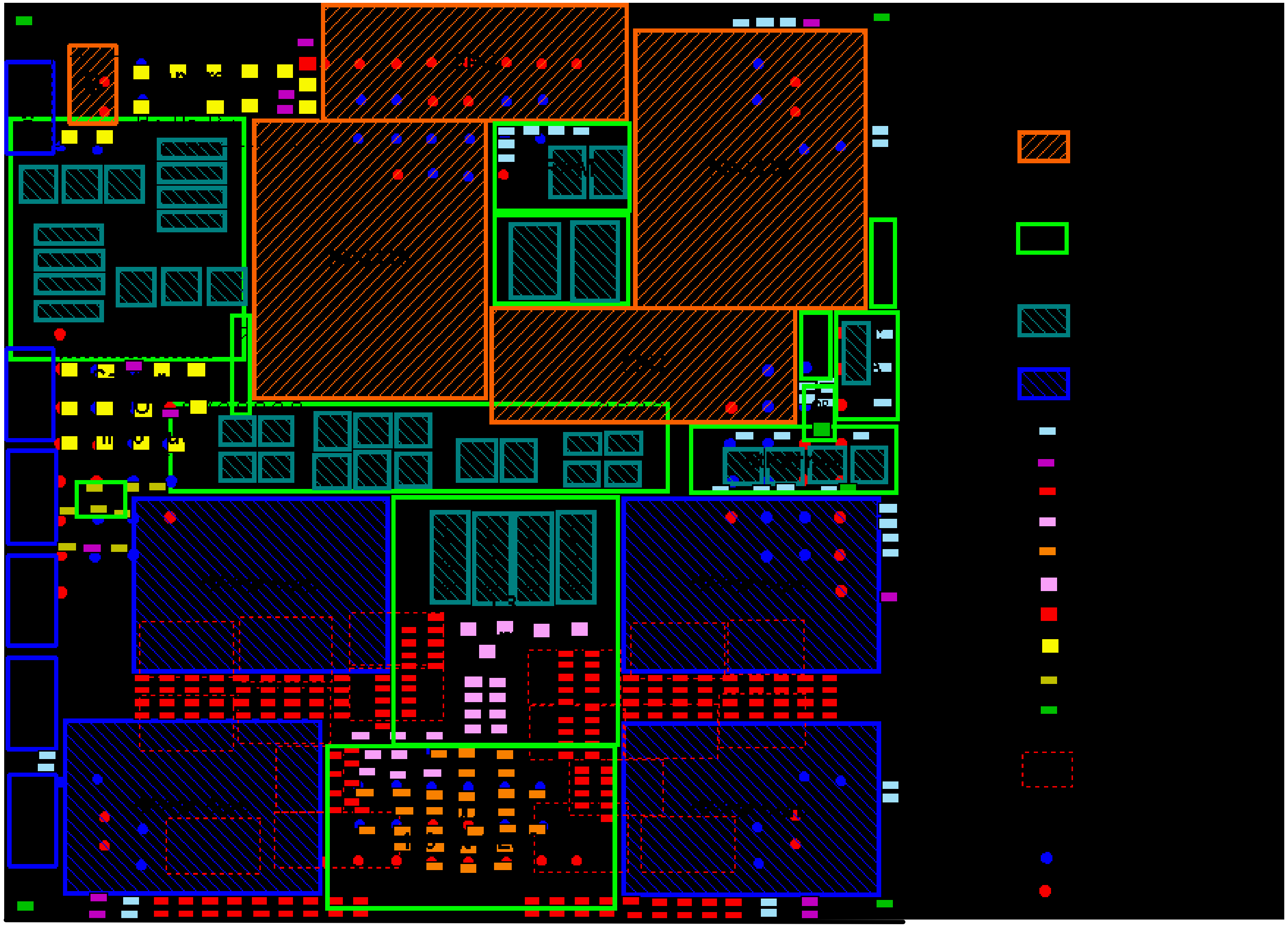
- File I/O and Host Interface



Control Network

- Boot, Monitoring and Diagnostics

Floor plan



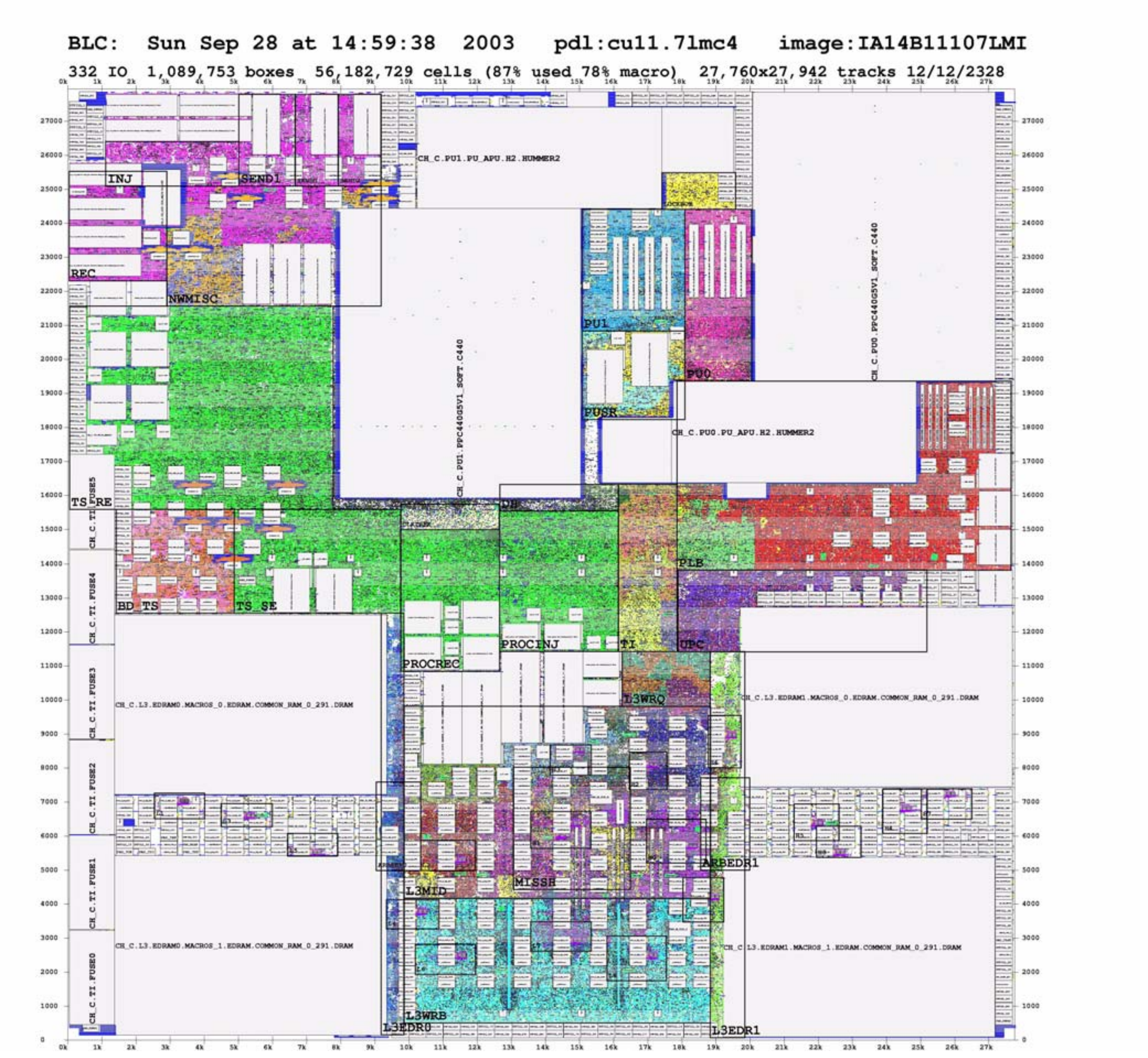
System-on-a-Chip

- IBM Cu-11 (0.13 mu technology) ASIC with:
  - hard cores -- dual (PPC440 + double FPU), PLL
  - soft cores -- Ethernet /DMA sub-system
  - custom I/O books for high speed signaling
  - eDRAM (32 Mb/chip)
  - SRAM (~2 Mb/chip), fuses, ECID

- “Custom design” twist: bitstacks
  - guided placement, auto wiring
  - critical for high speed send/capture of serialized Torus/Tree
  - far exceeds "normal" ASIC speeds -- up to 1.4 GHz clock.
  - Organizes wiring congestion at wide eDRAM ports

- IBM E&TS (Rochester, MN) style PD and test
  - careful clock design -- about 90 clock signals; 30 clock sub-domains
  - JTAG-based co-processor for in-system test/bring-up

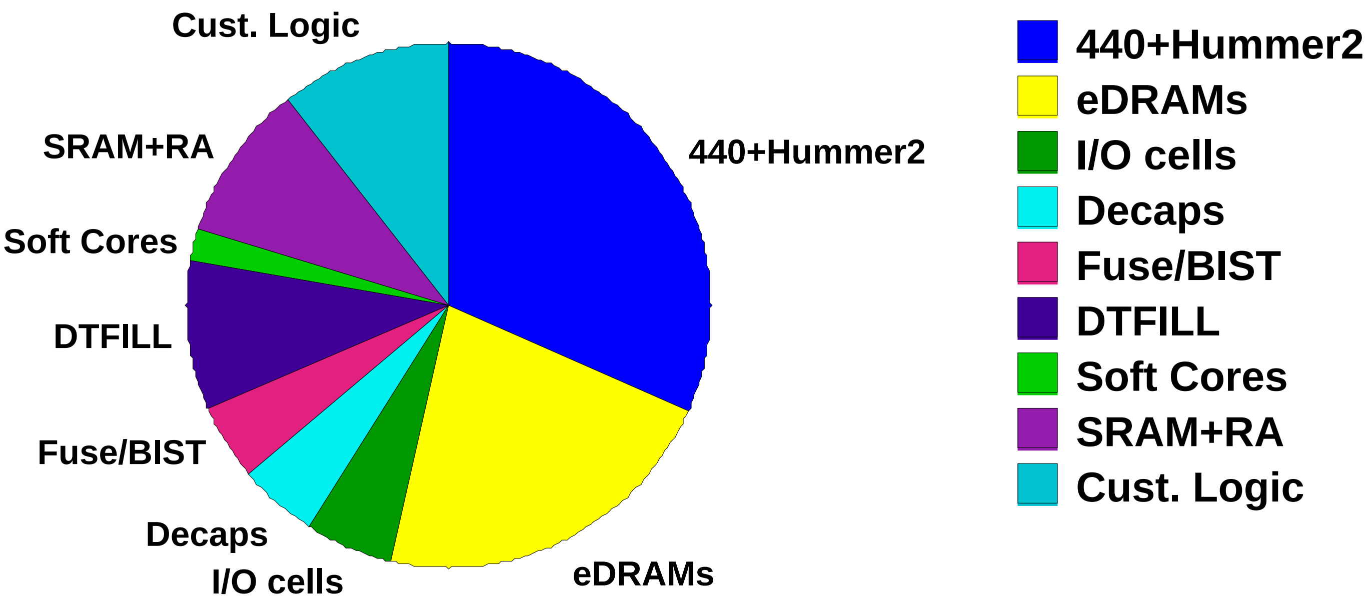
Physical Design



```
BLC:  Sun Sep 28 at 14:59:38  2003   pdl:cu11.7lmc4   image:IA14B11107LMI
332 IO  1,089,753 boxes  56,182,729 cells (87% used 78% macro)  27,760x27,942 tracks 12/12/2328
```

Chip area usage

Percent of chip



Prototype Bring-Up

- BLC DD1.0 Power-On on 06/16/2003

- Presently (10/10) > 600 chips running in various test stations
  - 1,2,8,32,128, 256, 512-ways
  - running anything from low level tests to applications to benchmarks
  - No show stoppers found.

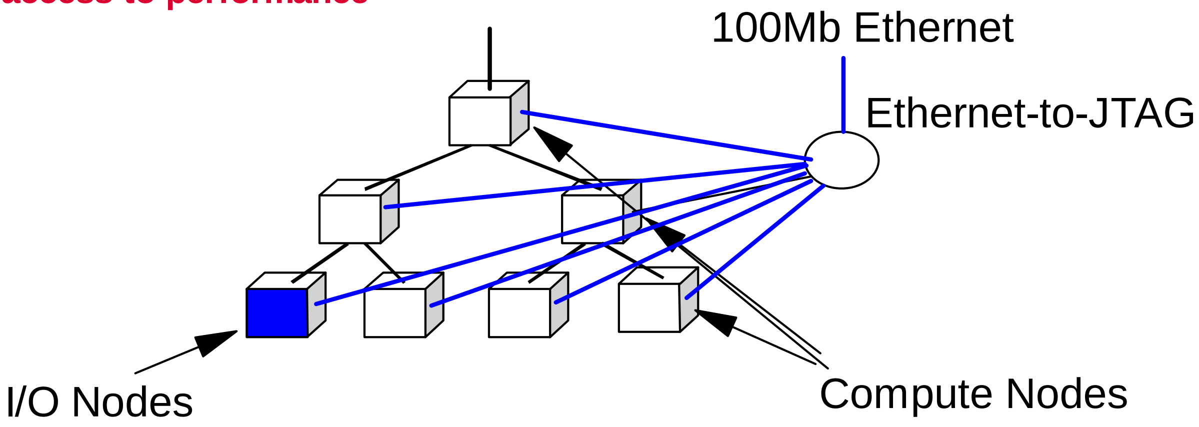
- The hardware works! Pace of bring-up limited by s/w and resources.
  - Outlook is good to do DD2.0 RIT in November

- BLC DD2.0 (production version):
  - No major functional difference
  - Better frequency
  - Improvements for robustness, diagnostics, error recovery.

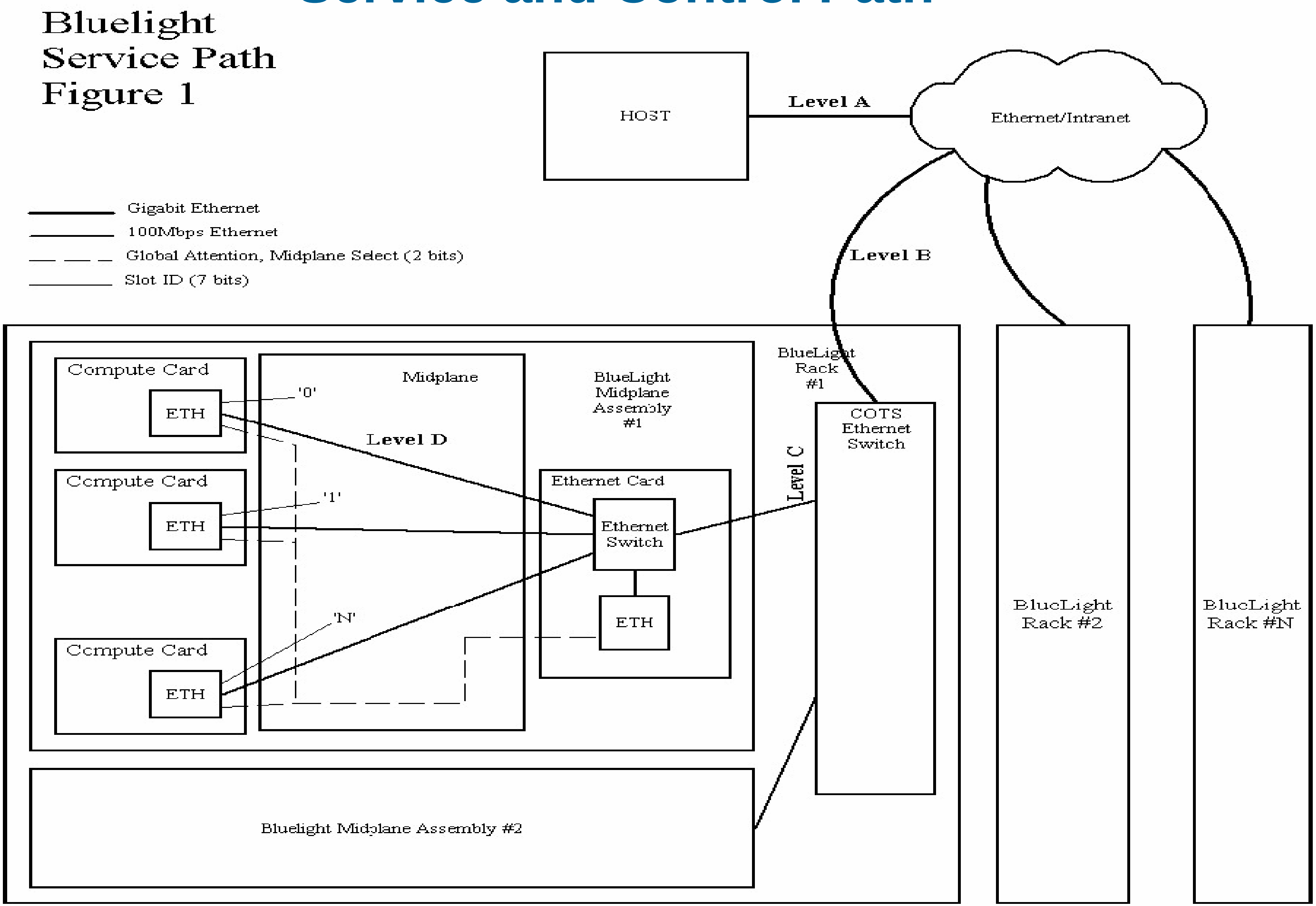
Control Network

Service Processor :
100Mb Ethernet to JTAG interface

- Direct access to any node
  - ➢ Partitioning
  - ➢ Configuration
- Direct access to shared SRAM in every node
  > boot-up code
  > messaging node <-> service processor
- In-system debug facilities
- Runtime noninvasive RAS support.
- Non-invasive access to performance counters



Service and Control Path



Bluelight Service Path Figure 1

Logic design

► Architect	Alan Gara
► Hummer2	Chuck Wait + team
► L2	Dirk Hoenicke, Martin Ohmacht
► L3, SRAM, Lockbox	Martin Ohmacht
► DDR controller	Jim Marcella
► Torus	Dong Chen, Pavlos Vranas, Sarabjeet Singh
► Tree	Dirk Hoenicke, Matt Blumrich
► High Speed Serial Comm	Alan Gara, Dong Chen, Sarabjeet Singh
► Global Interrupts	Dong Chen, Alan Gara
► EMAC4/PLB/DCR/BIC	Martin Ohmacht
► Performance Counters	Dirk Hoenicke
► Clock Tree	Matt Ellavsky
► Test & Bring-up structures	Marc Dombrowa, Ruud Haring, Steve Douskey, Mike Hamilton, Jim Marcella
► IOs	Ruud Haring
► Logic integration	Martin Ohmacht, Marc Dombrowa
► Libraries	Dan Beece

Logic verification

Team lead	Alan Gara
Stage releases	Martin Ohmacht
Model build, infrastructure	Dan Beece, Ruud Haring, Sandy Woodward + team
Regression	Lurng-Kuo Liu
Hummer2	Chuck Wait + team
Memory sub-system	Ben Nathanson, Brett Tremaine, Mike Wazlowski, Li Shang + designers Jim Goldade
Torus	Phil Heidelberger, Mike Wazlowski + designers
Tree	Burkhard Steinmacher, Brett Tremaine + designers
Tree Formal	Dirk Hoenicke, Steve German, Chris Zoellin
High Speed Serial Comm	Gerry Kopcsay, Minhua Lu
Global Interrupts	Lurng-Kuo Liu
Ethernet	Valentina Salapura, Jose Brunheroto
IPL & Bring-up	Marc Dombrowa, Ralph Bellofatto, Dong Chen, Martin Ohmacht
Test structures	Steve Douskey + team
Directed test cases	Krishna Desai + team (Bangalore)

Floor planning, synthesis, timing, DFT

► Team leads	Ruud Haring, Greg Ulsh, Fariba Kasemkhani
► Floor Planning	Terry Bright
► IO assignments	Ruud Haring
► Clocks	Matt Ellavsky
► Synthesis & pre-PD timing	Terry Bright, Jim Marcella, Chris Zoellin Dirk Hoenicke, Martin Ohmacht, Marc Dombrowa, Sarabjeet Singh
► Synthesis & timing support	Gay Eastman, Scott Mack + team
► Design for Testability	Marc Dombrowa
► AC test / AC-Lite	Ruud Haring

Physical Design

Team lead	Bob Lembach
Bit stacks	Terry Bright
Interface, ECO coordination	Jim Marcella
Clock Tree	Matt Ellavsky, Bruce Rudolph, Sean Evans
Physical Design	Mike Rohn, Cory Wood, Bruce Winter + others
post-PD timing closure	Jim Marcella, Todd Greenfield + team
Verity	Jim Marcella, Terry Bright

IBM Microelectronics

Interface	Scott Bancroft
CDS, RFQ	Terry Bright, Ruud Haring, Greg Ulsh, Paul Coteus, Mike Shapiro (IMD Austin)
AEs	Glen Smith (IMD RTP), Kurt Carlsen (IMD BtV)
.....	